Notations for High Efficiency Data Presentation in Mammography

Justin Starren, MD and Stephen M. Johnson, PhD

Department of Medical Informatics
Columbia University College of Physicians and Surgeons
New York, New York 10032

As a result of improvements in Medical Language Processing, the availability of categorical information (such as diagnoses or radiology findings) is increasing rapidly. This increased availability has created a need for more efficient methods for computer presentation. One method for developing such presentations would be to adapt the hand-written notation systems already used in paper-based records. We have characterized one such notation system, the Mammography Notation Sublanguage(MNS). The MNS is a true medical sublanguage with a definable lexicon and syntax. Compared with text reports, it represents a 37-fold size compression. A single "base" sublanguage pattern is identified for possible computer presentation of mammography findings. The issues involved in using such sublanguages for data presentation are discussed.

INTRODUCTION

Recent progress in Medical Language Processing has made it possible to automatically extract findings and diagnoses from radiology text reports[1]. As a result, there is a need for ways to present this extracted information to human users. Conventional full-text generation will be appropriate for some uses, but others, such as data validation or summary reports, will require more compact presentations. Methods have been developed for compressing large amounts of numeric data into small graphics [2,3]; there has been relatively little progress toward developing compact computer presentations of categorical data, such as radiologic findings.[4] In contrast to the situation with computers, paper medical charts already contain compact, efficient ways of presenting large amounts of categorical medical data. Physicians and other caregivers have developed compressed "shorthand" notations, in order to record data quickly and compactly. Common examples include "RRR" for "regular rate and rhythm" on a cardiac exam or the "stick figure" used to record pulses.

Adaptation of these paper-based notations could potentially allow much more efficient computer presentations of data. However, before such notations can be used for computer display of medical data, they must be defined with sufficient rigor that computer generation algorithms can be developed.

One method for developing such definitions is to treat the compressed notation system as a medical sublanguage and to leverage previous work in sublanguage analysis[5-8]. Sublanguage can be defined as:

> the particular language used in a body of texts dealing with a circumscribed subject ... in which the authors of the documents share a common vocabulary and common habits of word usage.[5]

Sublanguage analysis has been applied to several domains of medical narrative text including radiology reports and discharge summaries[5,6]. To our knowledge, formal sublanguage analysis of a medical notation system has not been reported.

As a domain for this analysis, we selected the compressed notation used by mammographers. For convenience we will call it the Mammography Notation Sublanguage (MNS). After a patient has a mammogram, the mammographer will review the films. Based on the findings, the mammographer may perform a physical exam or an ultrasound study. At the end, the mammographer will dictate a text report including all of the information and recommended follow-up (example in Table 1a). The mammographer also makes a notation of the results on the paper film jacket. This provides rapid reference for later mammographers as to how the studies were originally interpreted. If follow-up is needed, an entry will be made in a log book (Table 1b). When a biopsy is performed, the results are compared with the findings recorded in the log book.

0195—4210/96/$5.00

Because sublanguage analysis of a medical notation system has not been reported, this study has several goals: first, to generally characterize the MNS, as contrasted with free-text mammography reports; second, to demonstrate that the MNS is, in fact, a sublanguage; and last, to define a sublanguage grammar that could provide the basis of later computer data-presentation studies.

Table 1a - Text Report (body of report only)

DESCRIPTION:
The breasts are moderately dense which limits the sensitivy of mammography. The studies are compared with the previous studies of 1/1/93, 1/1/94 and 1/1/95. Nodular densities are noted bilaterally which may represent fluctuating cysts which was identified on previous studies.

There is a new cluster of microcalcifications in the upper outer quadrant of the left breast for which a needle localization and biopsy is recommended. These calcifications are indetermnate for malignancy and are new when compared with the prior studies.

IMPRESSION:
NEW CLUSTER OF MICROCALCIFICATIONS UPPER OUTER QUADRANT OF THE LEFT BREAST FOR WHICH A BIOPSY IS RECOMMENDED.

Table 1b - MNS Sample (log book entry)

new cluster μCA^{++} ⓁUOQ → Bx

METHODS

The hand-written log books for the Mammography Section of the Department of Radiology at the Columbia-Presbyterian Medical Center covering the period from 3/28/94 to 8/7/95 were transcribed into a computer readable format. Of the 353 log entries transcribed, 2 blank entries and 1 entry that did not refer to mammography results were discarded. In several cases, two distinct findings were listed in one entry; these were divided, resulting in 371 separate records. In roughly 5% of the records, the photocopy of the log book contained an illegible word. In these cases, the original log book and the full text report were reviewed. The length of mammography text reports was estimated by selecting 10 entries from the log book, manually retrieving the reports from the clinical information system, and performing word and character counts.

Table 2 - Classes and Subclasses in Mammography Notation

Class	Subclass	Definition
F		Finding
	F_m	Mammographic Finding
	F_u	Ultrasound Finding
L		Location
	L_a	Laterality
	L_r	Regionality
A		Attributes
	A_e	Physical Exam
	A_s	Size
	A_f	Form / Shape
	A_d	Distribution
	A_g	Grade / Severity
	A_n	Number / Count
	A_u	Ultrasound
	A_x	Attenuation (density)
T		Temporal Information
	T_d	Direction (new, increase, etc.)
	T_r	Referent (Since...)
R		Recommendation
	R_a	Additional Views
	R_s	Surgical Procedure
	R_u	Ultrasound
	R_b	Biopsy
D		Diagnosis

Each record was partitioned into a sequence of tokens, which represented terms. For example, "AD" and "arch. dist." are two different tokens, but both represent the term "Stromal Architectural Distortion". The number of tokens and terms in each record was recorded, as well as whether these were new, or had been encountered previously. The terms were manually inspected and assigned to classes and subclasses (Table 2). Terms involving clinical history, connectives and non-mammography terms were excluded from the encoding.

The sublanguage patterns were determined by converting each record into a pattern of subclass labels, and then progressively simplifying the patterns (Table 3). After the initial subclass encoding, the next step was to replace the subclasses in the pattern with their respective classes. In the example, $T_dA_dF_mL_aL_rR_b$ becomes **TAFLR.** Note that adjacent duplicates are reduced to a single instance: L_aL_r becomes **LL**, which becomes simply **L.** In order to better evaluate the semantic content of the

patterns, each pattern was reduced to a uniform order of classes. Specifically, **TAFLR** (the example) and **FLATR** both simplify to **AFLTR.** These most simplified patterns are termed "reduced".

Table 3 - Sample Encoding of Sublanguage Patterns

MNS	new cluster μCA^{++} ⓁUOQ → Bx
Subclass	$T_d A_d F_m L_a L_r R_b$
Class	T A F L R
Reduced	A F L T R

Sublanguage grammar analysis typically involves reduction of syntactic patterns into more standardized order prior to encoding[5,7]. Because of the desire to use the sublanguage grammar for data presentation, we delayed standardizing the order until the last step. An important measure of adequate sublanguage analysis is that the appearance of new patterns drops off as more entries are examined [7]. For convenience, we are calling this "saturation."

RESULTS

The MNS differs from free-text mammography reports in several ways. First is the high frequency of abbreviated tokens. In this corpus, 62% of the tokens are abbreviated. In text reports, appreviations occur rarely, if ever. Second is the occurrence of special symbols, such as Ⓡ for right or ↑ for increase. The most striking difference was the small size of the MNS records. The average record was only 19.6 characters long. In contrast, comparable text reports averaged 733.5 characters long, (even excluding demographic information and section labels). This represents a distillation of roughly 37-fold.

The MNS lexicon is also very compact. There were 161 terms, represented by 252 discrete tokens. Part of this compactness is due to the general paucity of articles, determiners, prepositions and other "helper" words. Medical reports in general are noted for their emphasis on noun phrases and relative paucity of verbs[8] In MNS, only one verb was present, "recommend" which was often represented as "→".

The terms clustered into 18 subclasses, which group into 6 classes (Table 2). The main class, *Finding,* is the primary abnormality (such as "mass"). *Finding* was the most common class represented, occurring in 98% of the records. Surprisingly, there were 7 entries with no finding, such as "new Ⓡ solid". For such records, the finding could be inferred from the attributes. In this case, the finding must be "mass" because only masses can be solid. *Location,* occurring in 84% of records, would include the side of the body and a region of the breast (e.g. "left upper outer quadrant"). *Attributes,* occurring in 71% of records, are the descriptors of the abnormality such as size or severity. MNS records are unusual in radiology because they include attributes relating to three different diagnositc modalities (x-ray mammography, ultrasound, and physical examination). In some records the diagnostic modality was listed along with the attribute (e.g. "US: solid"). But in the majority of cases, different types of attributes were mixed together (e.g. "solid spiculated palpable mass"). *Temporal* information (16%), *Recommendation* (12%), and *Diagnosis* (3%) were relatively rare.

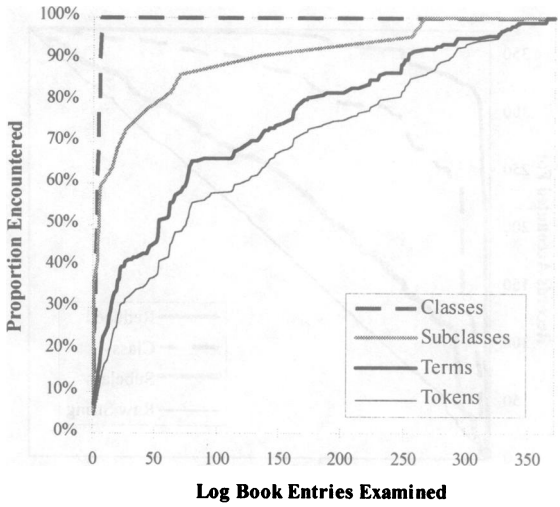

Figure 1 - Lexicon Saturation. *The entries were examined sequentially. Each line represents the total fraction of items (tokens, classes, etc.) encountered to that point. Once the graph reaches 100%, examination of more entries does not result in more items.*

Saturation was investigated in both the lexicon (Figure 1) and in sublanguage patterns (Figure 2). Athough the plots for neither tokens nor terms leveled off completely, the trend was clear. As expected, saturation at the subclass and class level was much more rapid. By entry 70 (19% of the corpus) 86% of the subclasses were encountered. By entry 6, all of the classes were represented.

Evaluation of the sublanguage patterns showed that, there were 339 unique strings, 235 distinct subclass patterns, 86 distinct class patterns, and 21 reduced patterns. The frequency distribution of class and

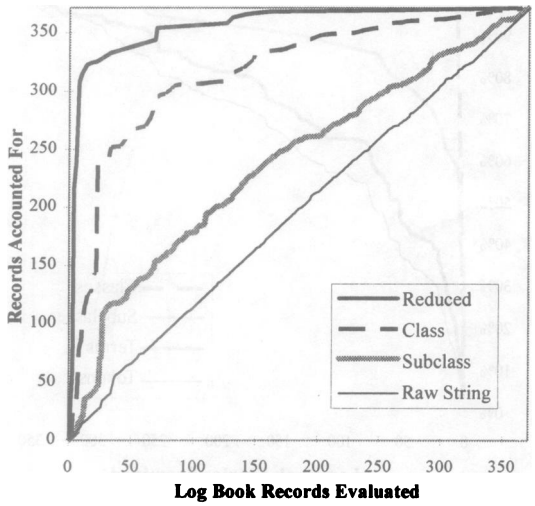

Figure 2 - Saturation of Patterns *The horizontal axis represents records evaluated sequentially. The vertical axis represents the aggregate number of records whose patterns have been encountered so far. (i.e. If the first pattern encountered is shared by 100 of the records, the graph value would be 100.)*

reduced patterns was not uniform. There were a few very common patterns and many rare patterns. The most common class pattern, **A F L**, occurred in 94 (25%) of the records. The top 5 class patterns accounted for half of the records. In contrast, 50 patterns appeared only once. The top 5 reduced patterns accounted for 304 (82%) of the records

The saturation of patterns is shown in Figure 2. Neither the raw strings nor the subclass patterns showed a significant tendency toward saturation. In contrast, class and reduced patterns showed evidence of saturation. By record 23 (6% of the corpus) class patterns accounting for 238 (64%) of the records were encountered. Reduced patterns accounting for 308 (83%) of the records were encountered by record 8 (2%). In spite of this, saturation did not appear complete, with new class and reduced patterns appearing up to record 361 (97% of the corpus).

The individual patterns were also inspected and pair-wise co-occurrance analysis performed to determine if a single fundamental pattern could be found which would account for the class ordering in all of the patterns observed. This was not the case. The four most common classes **(A,F,L,T)** occurred in all possible combinations. Even so, certain pairings were much more common. The base pattern (Table 4) accounts for over 57% of the records. Three closely related variants of this account for over 95% of the records.

Table 4 - Fundamental Patterns (all terms, except F, are optional and represent one or more instances)

Base Pattern	T_n A F L T_d D R
Variants	*A F A L A T D R*
	L_a A F A L_r A T D R
	T_n A F A L A T_d D R

DISCUSSION

The communication and data presentation needs within a medical subspecialty are very different from communication outside of the specialty or to the lay public. Compressed notations, like the MNS, evolved to fill this need. An important characteristic of these notions is that they assume an "expert audience". A significant portion of a mammography text report is to supply information to less expert persons who may read the report. In the example (Table 1a), the breast was dense (a common benign condition). The mammographer used the sentence "The breasts are moderately dense *which limits the sensitivity of mammography.*" This comment was added for other caregivers. Every mammographer knows the effect of breast density on sensitivity; this is added for non-mammographers.

Analysis of the MNS demonstrates that it does have the characteristics of a true sublanguage. This is clearly seen at the class level in both the lexicon and in the sublanguage patterns. The classes and common patterns are encountered very rapidly. The continued occurrence of rare patterns is not inconsistent with the general closure of the sublanguage. As a medical specialty changes, the sublanguage of that specialty must change in order to stay relevant. In fact, changes in sublanguage have been suggested as one way to track changes in a scientific field[7].

The underlying motivation of this work is to use the MNS for computer display of mammography findings. Several characteristics of the MNS have implications for its use in computer displays. The most important is its compact size. Typically, a mammogram report requires numerous lines of a computer screen for display. In contrast, MNS provides a way to present results in one or two lines. Figure 3 shows a hypothetical summary display using MNS.

A computer adaptation of "pure" MNS will be likely be suitable for displays geared toward mammographers. Our plan is to use it for displaying the results of our Medical Language Processing system.[1] The use of MNS need not be limited to

mammographers. Anecdotally, many of the special symbols, including laterality (Ⓡ Ⓛ), change (↑ ↓), implication (→), positive/negative(⊕ ⊖) and terms based on Latin(p̄ c̄), appear in the notations of other specialties.

The development of these non-ASCII symbols is probably related to the hand-written nature of the MNS. The special symbols reduce ambiguity without increasing space. The letter "L" might mean "left" or "lower", but the symbol Ⓛ is unambiguous and is much faster to write (and possibly read) than is "left". Even if a computer display is unable to support the special symbols, the grammar of MNS could be used with text expansions of the terms to create a "telegraphic" style of display. In addition, this "written out" form of MNS may be well suited to the display of mammography results to non-mammographers who may be familiar with the terminology but not the symbolology. The example, **new μCA^{++} ⓁUOQ → Bx**, would expand to "*new cluster microcalcifications left breast upper outer quadrant, recommend biopsy*". This expanded form would be intelligible to nearly all caregivers.

Exam	Date	Results
Ⓑ Mam	3/4/91	FA Ⓡ LIQ o/w ⊖
Ⓑ Mam	4/8/92	⊖△
Ⓑ Mam	3/12/93	⊖△
Ⓑ Mam	4/10/94	new μCA^{++} ⓁUOQ → Bx
Ⓛ NL	4/22/94	⊕ loc
Ⓑ Mam	5/6/95	⊖

Figure 4 - Possible Summary Display Using MNS

CONCLUSION

In summary, the MNS represents a distinct medical sublanguage with a defined lexicon and syntax. Computer data presentations based on notational sublanguages, like MSN, have the potential to provide efficient, compact display of very complex information. We will be implementing a computer data presentation based on MNS and evaluating its usefulness at displaying mammography findings to mammographers. We also hope to extend this work to other medical notation systems.

Acknowledgment

We would like to thank Dr. Suzanne Smith, for access to the Mammography Logs and for help in interpreting the entries. This publication was supported in part by grant LM07079 from the National Library of Medicine and by the New York State Science and Technology Foundation.

References

1. Friedman C, Hripcsak G, DuMouchel W, Johnson SB, Clayton PD. Natural Language Processing in an Operational Clinical Information System. *Natural Language Engineering* 1995;1(1):83-108.
2. Powsner SM, Tufte ER. Graphical summary of patient status. *Lancet* 1994;334:386-9.
3. Cole WG, Stewart JG. Human performance evaluation of a metaphor graphic display for respiratory data. *Methods of Information in Medicine* 1994;33(4):390-6.
4. Preiss B, Kaltenbach M, Zanazaka J, Echave V. Concept graphics: a language for medical knowledge. *Proceedings - the Annual Symposium on Computer Applications in Medical Care* 1992;515-9.
5. Hirschman L, Sager N. Kittredge R, Lehrberger J, editors. *Sublanguage. Studies of Language in Restricted Semantic Domains*. Berlin: Walter de Gruyer; 1982; 2, Automatic Information Formatting of a Medical Sublanguage. p. 27-80.
6. Sager N, Lyman M, Bucknall C, Nhan N, Tick LJ. Natural language processing and the representation of clinical data. [Review]. *JAMIA* 1994;1(2):142-60.
7. Harris Z. *A theory of Language and Information. A Mathematical Approach.* Oxford: Clarendon Press; 1991.
8. Sager N. Kittredge R, Lehrberger J, editors. *Sublanguage. Studies of Language in Restricted Semantic Domains*. Berlin: Walter de Gruyer; 1982; 1, Syntactic Formatting of Science Information. p. 9-26.